Exquisitor at the Lifelog Search Challenge 2019

Omar Shahbaz Khan
IT University of Copenhagen
Copenhagen, Denmark
omsh@itu.dk

Björn Þór Jónsson
IT University of Copenhagen
Copenhagen, Denmark
bjorn@itu.dk

Jan Zahálka
bohem.ai
Prague, Czech Republic
jan.zahalka@bohem.ai

Stevan Rudinac
University of Amsterdam
Amsterdam, Netherlands
s.rudinac@uva.nl

Marcel Worring
University of Amsterdam
Amsterdam, Netherlands
m.worring@uva.nl

ABSTRACT

Interactive learning is an umbrella term for methods that attempt to understand the information need of the user and formulate queries that satisfy that information need. We propose to apply the state of the art in interactive multimodal learning to visual lifelog exploration and search, using the Exquisitor system. Exquisitor is a highly scalable interactive learning system, which uses semantic features extracted from visual content and text to suggest relevant media items to the user, based on user relevance feedback on previously suggested items. Findings from our initial experiments indicate that interactive multimodal learning will likely work well for some LSC tasks, but also suggest some potential enhancements.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Multimedia databases**.

KEYWORDS

Lifelogging; Interactive learning; Exquisitor.

ACM Reference Format:
Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2019. Exquisitor at the Lifelog Search Challenge 2019. In *Lifelog Search Challenge'19 (LSC'19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3326460.3329156

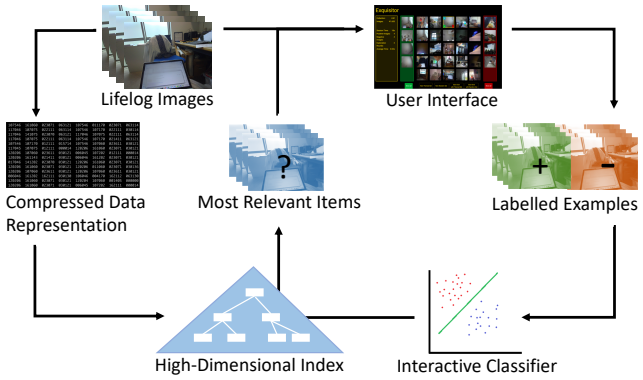

Figure 1: Exquisitor's interactive learning pipeline. Initially, the lifelog's image collection is processed to produce a compressed semantic representation, that is stored in a scalable high-dimensional index. In each round of the interactive learning process, the user is shown a set of potentially relevant images. The user's judgments are then used to train a classifier, which in turn is used to retrieve a new set of images to show to the user. With the LSC collection, producing new suggestions takes about 30ms on a laptop computer.

1 INTRODUCTION

Today's plethora of small devices allows capturing a tremendous amount of personal information. The people who make use of these devices to the fullest extent, gathering a variety of information about their daily lives, are termed lifeloggers. The most important feature of a typical *lifelog* is the image collection generated by a camera attached to the individual lifelogger taking pictures at regular intervals. The lifelog can also contain other sensor data, such as temperature, location, heart rate, and audio, depending on which devices the individual uses. Furthermore, this data can be processed with state-of-the-art computer vision and learning algorithms to produce semantic annotations. Applications of such personal lifelog data include self-monitoring and assisted memory [10].

The Lifelog Search Challenge (LSC) is a competition where researchers are asked to study and develop methods to solve search-related tasks for a multimodal lifelog dataset. Each task in LSC is an independent query, to be solved in a few minutes, where a correct result is a single image returned from a set of relevant images. The query description is given gradually, as might be typical when a lifelog is used to find information and the user slowly remembers more details about the situation. The first edition of LSC, held in 2018, showcased a variety of multimedia browsers aiming to search the lifelog with different approaches, ranging from traditional keyword search to novel virtual reality-based approaches [8].

Working with a lifelog should be a highly interactive process, where the lifelog user is collaborating with the lifelog system on a variety of tasks, ranging from pure exploration of the lifelog collection to focused search tasks to retrieve images relating to

LSC'19, June 10–13, 2019, Ottawa, ON, Canada

ACM ISBN 978-1-4503-6781-3/19/06…$15.00
https://doi.org/10.1145/3326460.3329156

particular memories. Multimedia analytics has been proposed as a research area aimed exactly at solving such diverse interactive information needs [23]. In multimedia analytics, an analytical session is composed of multiple different sub-tasks, ranging from browsing to seeking a particular known item, thus forming an exploration-search axis. Furthermore, *interactive multimodal learning* was proposed as an umbrella task capable of satisfying all the tasks on the exploration-search axis [23]. It is therefore of significant interest to apply interactive multimodal learning to LSC.

We have recently developed Exquisitor, a highly scalable interactive multimodal learning approach [13]. Figure 1 illustrates the iterative feedback process employed by Exquisitor as employed with lifelog data. When a lifelog user has an information need, she is initially presented with a set of randomly selected images from the lifelog and asked to give feedback on (some of) the items. The feedback is used to build (and subsequently update) a classification model, which in turn is used to provide new suggestions; this iterative process continues as long as the user deems necessary. A key feature that sets Exquisitor apart from other interactive learning approaches is its scalability: Exquisitor can retrieve suggestions from the YFCC100M collection with sub-second latency, using computing resources that are comparable to today's high-end mobile device. In this paper, we propose to use Exquisitor to solve the tasks of the Lifelog Search Challenge.

The remainder of the paper is organized as follows. In Section 2, we briefly give background for interactive learning and LSC. Section 3 then outlines the Exquisitor approach and its exploration interface. In Section 4, we look at the dataset provided by LSC and describe the processing required to use it with the Exquisitor approach. In Section 5, we briefly report on initial experiments with interactive retrieval tasks, before concluding the paper in Section 6.

2 BACKGROUND

Interactive learning comes in two basic forms, *active learning* and *user relevance feedback* [11]. In active learning, the goal is to create the best possible classifier, so the contribution of the user is typically to annotate samples close to the decision boundary between classes [2, 12]. User relevance feedback algorithms, in contrast, focus on giving users insight into the multimedia collections [17]. As a result, relevance feedback systems typically present as suggestions to the user the items for which the classification model is the most confident [19]. While this latter strategy may require more interactions to achieve the same final quality of the classification model, users may achieve their desired knowledge earlier [23].

Originally proposed in the 90s, early user relevance feedback systems for content-based image and video retrieval commonly relied on visual features that lack meaningful representation, such as colour, texture, shape and edge histograms [19], as well as indexing techniques that are inefficient in high-dimensional spaces, such as R-trees and kd-trees [4]. While relatively little work has been done on user relevance feedback in the last decade, recent advances in both high-dimensional indexing and data representation, along with calls for action from the multimedia community [21, 23], have motivated us to re-visit user relevance feedback with the Exquisitor approach [13].

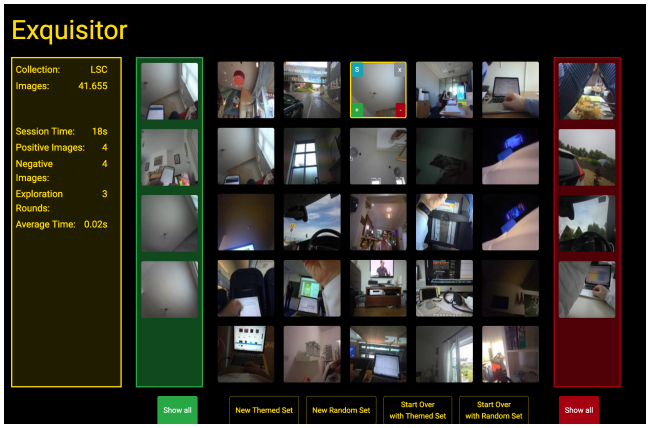

Figure 2: Exquisitor's browser-based user interface. When hovering over an image, the user can label it as positive (bottom left), negative (bottom right), or seen (top right). Positive items (green column) and negative items (red column) are then used for updating the model.

Lifelogging is also steeped in history. In 1945, Vannevar Bush published an article in which he proposed the "Memex", which he described as "a device capable of storing all books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility" [1]. Despite the desire for such a device, the required technology did not exist and therefore Bush could only encourage future researchers to carry out this vision. The pioneering effort was the MyLifeBits project [5], where Gordon Bell attempted to digitize nearly every aspect of his life, creating the first lifelog. Recent years have seen the emergence of more devices capable of capturing lifelog data, such as miniature cameras, heart rate monitors, audio capturing devices, and GPSs, to name a few. Collecting all this information is the first step, of course, but as with the MyLifeBits project, the ability to process the data in real time at scale in a flexible way is still desired.

The LSC, now in its second year, is the first interactive challenge focusing on lifelog data. It derives its format from VideOlympics [22] and Video Browser Showdown (VBS) [20], inviting interactive retrieval systems to solve interactive tasks at premise. Six teams participated in LSC 2018. Some of these had previously participated in VBS, while others were new systems; overall the more developed systems had greater success [8]. The techniques of the different retrieval systems varied significantly, but features such as filters, similarity search and keyword search were a recurring theme [8]. On top of these, specific systems emphasized different interactions, such as virtual reality [3], sketch-based [14, 15] or visual concepts [16] to name a few. However, none of the LSC 2018 participants used a relevance feedback-based approach.

3 EXQUISITOR

Exquisitor is a user relevance feedback approach capable of handling large scale collections in real time [13]. It uses a Linear SVM classifier as the underlying model deployed to score items in a compressed feature space each interaction round. Furthermore it uses a high dimensional indexing approach based on extended Cluster

Pruning (eCP) [6]. The Exquisitor system used for LSC consists of three parts: (1) a web-based user interface for receiving and judging submissions; (2) an interactive learning server, which receives user judgments and produces a list of suggestions; and (3) an off-the-shelf web server which serves image thumbnails. In the following, we describe the first two parts of the system.

3.1 Exquisitor Interface

The current Exquisitor user interface, shown in Figure 2, is browser-based. It is largely a traditional interactive learning interface, in that users are asked to label positive and negative examples, which are then used to learn their preferences and determine the new round of suggestions. Due to the extreme efficiency of the scoring process, however, the interface itself initiates the request for new suggestions, either at regular intervals or when new examples have been produced.

3.2 Exquisitor Server

Exquisitor is developed to handle large-scale image collections, where each image is described with feature vector data from the visual and text modalities. The main components of the system are a) data representation and indexing, and b) the scoring process. We will briefly describe these in the following.

The high-dimensional feature vectors from the visual and text modality are independently compressed using an index-based compression method [24], where each feature vector is represented using the top 6 features of each modality and compressed into only three 64-bit integers. This results in an item only requiring 24 bytes of space per feature vector modality. The system has no need for decompression as it is capable of scoring the items directly in compressed space.

The compressed feature vectors are then indexed using the eCP high-dimensional indexing algorithm [7]. A set of R representative vectors is chosen from the collection and each vector is assigned to the closest representative, thus forming clusters in the compressed high-dimensional space. To facilitate retrieval, the clusters are recursively indexed, using the same method to select representatives of the representatives, to a chosen height L of the index.

In each interaction round, the Linear SVM model yields a classification hyperplane, which is used to form a farthest neighbor query to the cluster-based index. The goal is to yield k suggestions, which can be presented to the user. From each modality, b clusters are retrieved and their contents scanned to yield the r furthest neighbors from hyperplane. Using late modality fusion, these r candidates from each modality are then merged with a rank aggregation scheme to produce one ranked list, and the top k overall candidates returned. If further efficiency is required, multiple CPU cores can collaborate in producing the answer, by using w workers to process b/w clusters each.

Table 1 summarizes the initial parameter settings we have used for the LSC collection. Note that experiments with YFCC100M have shown that there is a tradeoff between latency and result quality. As more clusters are processed (higher b) both latency and result quality increase, but at some point result quality stops improving, and may even get worse with additional processing in some cases. We have yet to determine the optimal tradeoff between latency and result quality for the LSC data, but the collection is small enough to fully process in about 20 milliseconds per interaction round using only a single CPU core.

Table 1: Runtime parameters for Exquisitor with LSC data.

Parameter	Description	Default
Offline Indexing Parameters		
R	Number of representatives/clusters	417
L	Height of index tree	2
Runtime Scoring Parameters		
b	Clusters read from the index	16
r	Candidate items from each cluster	100
k	Number of new suggestions returned	25
w	Number of CPU cores used	1

4 DATASET PREPARATION

LSC 2019 provides a dataset consisting of lifelog data collected from a single user over the course of 27 days [9]. The dataset consists of 41.665 images with associated metadata and biometric data of the lifelogger. This section describes how the given data was processed into visual and textual feature vectors for use with Exquisitor.

4.1 Visual Data

In the LSC dataset, visual concepts (e.g., "computer", "indoor", and "wall") have already been assigned to images with a certainty score ranging from 0 to 1. All in all, there are 548 unique concepts in the collection; the highest number of concepts found on a single image is 15. As described above, the 6 visual concepts with the highest certainty scores are retained in the compressed data representation, while the remaining concepts are ignored.

Note that not all images have visual concept data. A total of 986 images have apparently not successfully cleared the concept generation process and are not represented in the visual dataset at all. Additionally, 1,454 images had a "null" concept assignment, indicating that the feature extraction process yielded no concepts. As the data was processed sequentially, according to time and date, we have made the assumption that most images with missing visual concept data can be represented by the features of the previous successful image. Figure 3 shows one example where this assumption holds, but there are also examples where the previous image is far less similar.

Figure 3: An example of consecutive images from the LSC dataset, where the first image has valid visual concepts while the following images have none.

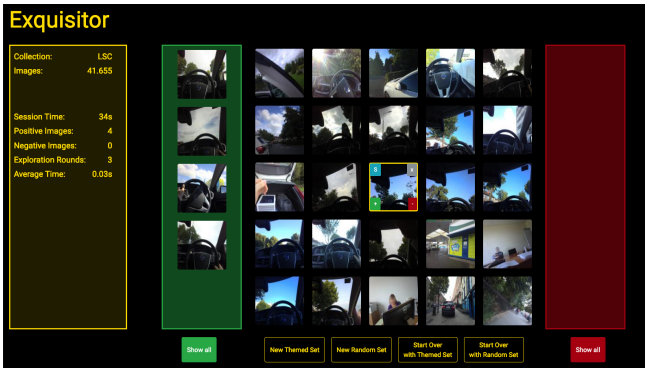

Figure 4: Suggestions provided in the third round of interaction for an example AS task: find images of driving.

4.2 Textual Data

The text metadata consists of annotated descriptions for 23,788 images. Along with a general description of the activity in the image, the direct object with which the user was interacting, if any, is also given as text. These two fields were used to extract text feature scores, using a 100-topic LDA model trained on the English Wikipedia corpus using the gensim toolkit [18]. Note that 409 of the 986 items that had no visual features were found to have some textual features, leaving 577 images without any visual or text features.

4.3 Other Data

Additional metadata about the lifelog user were provided, such as location, heart-rate, food information, etc. For now, these have not been used to extract features, but this remains an option. Furthermore, this information could be used to combine filters with the relevance feedback process.

5 INITIAL EXPERIENCES

According to [23], the way a user initially interacts with a collection is by browsing through it. As further insight is gained about the collection and the task of the user becomes more clear, the user can start to narrow the scope until a result is achieved. The question then is whether this type of process is suitable for LSC tasks.

Based on the previous Lifelog Search Challenge, and other similar competitions such as Video Browser Showdown, the likely tasks for systems can be categorized into two groups: *Known Item Search* (KIS) and *Ad-hoc Search* (AS). In LSC 2018, the former was dominant. KIS tasks means that there is only one image (or a small set) that will satisfy the query, while AS tasks have more broad answers. In the following, we consider examples of AS and KIS tasks, and describe our initial experiences.

5.1 Example: Ad-hoc Search

As an example AS task, consider finding images where the user is driving. As can be seen in Figure 4, it took only 3 interaction rounds, starting from a random set of images, before the system became well aware of our intent and provided many relevant results. The overall process took a little over 30 seconds (left column), while producing suggestions took on average 30 milliseconds per interaction round.

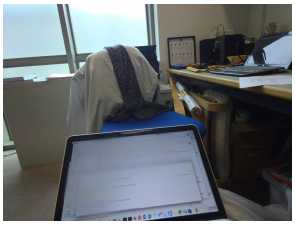

Figure 5: Random image chosen for the example KIS task.

5.2 Example: Known Item Search

We believe that a KIS task will be harder with a relevance feedback-based system, as finding a suitable Linear SVM model that separates the correct image from the collection will be hard. To test this, we have randomly chosen the image in Figure 5 as an example of a KIS task. Starting again from a random set of images, Exquisitor quickly identified that the information need included laptops or computers, but as the image is very similar to many other images containing laptops or computers, the correct image could not be found in 40 interaction rounds. Note that in LSC, each task generally has a set of images considered relevant, so this example KIS task is most likely significantly harder than LSC tasks, but it was nevertheless instructive, as summarized next.

5.3 Summary of Observations

So far, our work has been more focused on exploration than on identifying known items. While relevance feedback alone should be capable of narrowing the scope of exploration and eventually finding the correct items, some additional functionality appears necessary for the time-constrained LSC tasks. We have identified the following key issues to address before LSC 2019 starts:

(1) Using more modalities than only visual and text modalities is the first priority. Metadata, such as location, time, and day, could be used both to find candidates and influence their ranking, thus impacting the choice of suggestions. Furthermore, filters on metadata could be used to reduce the scope of exploration, thus allowing users to more quickly arrive at a correct answer.
(2) Currently, the initial set of images is chosen randomly. Using either a visual query or text query to prime the suggestions could be a good addition to the interface. Due to the underlying index structure, such queries can be easily implemented without changing the relevance feedback process.
(3) When looking for a known item, it must be possible to instruct the system that, while all of the suggestions shown are indeed relevant, none of them are *exactly* what is sought. In that case, the system should show further suggestions based on the same model.
(4) Currently, the interface only shows image thumbnails. Examining an image in more detail, along with its metadata, could help the user evaluate its relevance, and potentially also help choose which modalities to use or to adjust filters.

6 CONCLUSION

In this paper we have described the initial configuration of the Exquisitor system for our first participation in the Lifelog Search Challenge (LSC 2019). Exquisitor is a highly scalable interactive learning system, which relies on user relevance feedback to improve its model of the user's information need. What sets this system apart from related work is the scalability, which it owes to innovative feature selection, compression and indexing as well as the ability to train the interactive model and score multimedia items directly in the compressed space. As a consequence, the visual and text features for the LSC collection can be stored in less than 6MB of RAM and processed in about 30 milliseconds on average per interaction round on a modest laptop computer. We have described our initial experiences with using Exquisitor on lifelog data, and proposed a number of enhancements to the system for improved performance.

REFERENCES

[1] Vannevar Bush. 1945. As We May Think. *The Atlantic Monthly* 176, 1 (1945), 101–108.

[2] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active Learning with Statistical Models. *J. Artificial Intelligence Research* 4, 1 (1996), 129–145.

[3] Aaron Duane, Cathal Gurrin, and Wolfgang Huerst. 2018. Virtual Reality Lifelog Explorer: Lifelog Search Challenge at ACM ICMR 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*. ACM, Yokohama, Japan, 20–23.

[4] Myron Flickner, Harpreet S. Sawhney, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. 1995. Query by Image and Video Content: The QBIC System. *IEEE Computer* 28, 9 (1995), 23–32.

[5] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. 2002. MyLifeBits: Fulfilling the Memex Vision. In *Proc. ACM Multimedia*. ACM, Juan les Pins, France, 235–238.

[6] Gylfi Þór Guðmundsson, Laurent Amsaleg, and Björn Þór Jónsson. 2012. Impact of Storage Technology on the Efficiency of Cluster-based High-dimensional Index Creation. In *Proc. International Conference on Database Systems for Advanced Applications (DASFAA)*. Springer, Busan, South Korea, 53–64.

[7] Gylfi Þór Guðmundsson, Björn Þór Jónsson, and Laurent Amsaleg. 2010. A Large-scale Performance Study of Cluster-based High-dimensional Indexing. In *Proc. International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval (VLS-MCMR)*. ACM, Firenze, Italy, 31–36.

[8] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, et al. 2019. Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59.

[9] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Münzer, Rami Albatal, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. 2019. A Test Collection for Interactive Lifelog Retrieval. In *Proc. International Conference on MultiMedia Modeling (MMM)*. Springer, Thessaloniki, Greece, 312–324.

[10] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends in Information Retrieval* 8, 1 (2014), 1–125.

[11] Thomas S. Huang, Charlie K. Dagli, Shyamsundar Rajaram, Edward Y. Chang, Michael I. Mandel, Graham E. Poliner, and Daniel P. W. Ellis. 2008. Active Learning for Interactive Multimedia Retrieval. *Proc. IEEE* 96, 4 (2008), 648–667.

[12] Miriam W. Huijser and Jan C. van Gemert. 2017. Active Decision Boundary Annotation with Deep Generative Models. In *Proc. IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Venice, Italy, 5296–5305.

[13] Björn Þór Jónsson, Omar Shahbaz Khan, Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Jan Zahálka, Stevan Rudinac, Gylfi Þór Guðmundsson, Laurent Amsaleg, and Marcel Worring. 2019. Exquisitor: Interactive Learning at Large. arXiv:1904.08689.

[14] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. VIRET: A Video Retrieval Tool for Interactive Known-Item Search. In *Proc. ACM International Conference on Multimedia Retrieval (ICMR)*. ACM, Ottawa, ON, Canada.

[15] Jakub Lokoč, Tomáš Souček, and Gregor Kovalčík. 2018. Using an interactive video retrieval tool for lifelog data. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*. ACM, Yokohama, Japan, 15–19.

[16] Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, Manfred Jürgen Primus, and Klaus Schoeffmann. 2018. lifeXplore at the Lifelog Search Challenge 2018. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*. ACM, Yokohama, Japan, 3–8.

[17] Chris North. 2006. Toward Measuring Visualization Insight. *IEEE Computer Graphics and Applications* 26, 3 (2006), 6–9.

[18] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[19] Yong Rui, Thomas S. Huang, and Sharad Mehrotra. 1997. Content-Based Image Retrieval with Relevance Feedback in MARS. In *Proc. International Conference on Image Processing (ICIP)*. IEEE Computer Society, Santa Barbara, CA, USA, 815–818.

[20] Klaus Schoeffmann. 2014. A User-Centric Media Retrieval Competition: The Video Browser Showdown 2012-2014. *IEEE MultiMedia* 21, 4 (2014), 8–13.

[21] Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, George Awad, and Jakub Lokoč. 2018. Interactive Video Search: Where is the User in the Age of Deep Learning?. In *Proc. ACM Multimedia*. ACM, Seoul, Republic of Korea, 2101–2103.

[22] Cees G. M. Snoek, Marcel Worring, Ork de Rooij, Koen E. A. van de Sande, Rong Yan, and Alexander G. Hauptmann. 2008. VideOlympics: Real-Time Evaluation of Multimedia Retrieval Systems. *IEEE MultiMedia* 15, 1 (2008), 86–91.

[23] Jan Zahálka and Marcel Worring. 2014. Towards Interactive, Intelligent, and Integrated Multimedia Analytics. In *Proc. of the IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE Computer Society, Paris, France, 3–12.

[24] Jan Zahálka, Stevan Rudinac, Björn Þór Jónsson, Dennis C. Koelma, and Marcel Worring. 2018. Blackthorn: Large-Scale Interactive Multimodal Learning. *IEEE Transactions on Multimedia* 20, 3 (2018), 687–698.